Streamlining the CERIF XML Data Exchange Format Towards CERIF 2.0

Brigitte Jörg[a b], Jan Dvořák[c], Thomas Vestdam[d]
[a] German Research Center for Artificial Intelligence (DFKI GmbH), Germany
[b] euroCRIS, The Netherlands
[c] Institute of Information Studies and Librarianship, Faculty of Arts, Charles University in Prague, Czech Republic
[d] Atira A/S, Denmark

Summary

The Common European Research Information Format (CERIF) is an established standard for Current Research Information Systems (CRISs). CRISs face the increasing need for information sharing and exchange. euroCRIS released the first official CERIF XML exchange format in 2007; it followed the structure of the relational data model. Based on experience with the format and consulting with the CRIS community on newer interoperation and exchange concepts, the authors proposed an update to the CERIF XML exchange format. This updated CERIF XML aimed at compactness with expression and backwards compatibility. This article provides insight into the motivation for change, presents the updated format, and finally outlines possible next steps.

1 Introduction

Research Information Systems (CRISs) provide multiple stakeholders with the required data for their activities and have been recognized in playing an increasingly critical role with information sharing (Ivanovic 2011, Jeffery and Asserson 2010, van Godtsenhoven et. al. 2009, Hornbostel 2006, Zimmermann 2002, Jeffery et al. 1989). International initiatives aiming at large scale interoperability of scientific information date back to the Seventies, where a UNESCO/ICSU committee was set up to investigate the feasibility of a world science information system. A then published report by UNISIST (Martin 1974) concluded, that such a system is both "necessary and feasible" (Wysocki and Tocatlian 1979). CERIF[1] as a standard format and EU recommendation to Member States arose from a Conference of European University Rectors and in parallel from the heads of research funding organizations of G7 countries (Jeffery 1999). In 2002, the European Commission entrusted euroCRIS to take care of CERIF, which has since been improved and continuously updated; increasingly in tight cooperation with vendors and suppliers of CRIS systems and alongside research activities i.e., real world requirements. For historical reasons,

[1] The Common European Research Information Format (CERIF); a EU Recommendation to Member States: http://cordis.europa.eu/cerif/, http://www.euroCRIS.org/

CERIF has been developed and described in relational terms; the first official CERIF XML release (Joerg et al. 2007) was a direct 1:1 transformation of this underlying ER-Model and has since been maintained in relational-like structure and inline with on-going CERIF model updates. However, novel CRIS interoperation and information exchange concepts must consider a wider scope, and allow for more independence from underlying system structures, because fully-fledged or *ideal* CERIF is not always a realistic option. In fact, such a scenario applies to any existing *non*-CRIS, where CERIF XML is still relevant for exchange, but where mappings or transformations to an ER-centric XML structure become difficult. This is regardless to whether a given system is actually able to represent exactly the same *information* as is possible with ideal CERIF or not. Such problems are often said to arise from an *impedance mismatch* between different representations and pose well-known issues with transformations between object-oriented and relational models, as well as between the system analysis and design (Castro et al. 2002). Within the CRIS community the issues have been timely recognized and addressed. While the 2007 CERIF XML exchange format was functional and could express any CERIF data, it had certain aspects that made it difficult to work with for data exchange.

The updated CERIF XML exchange format aimed at removing the rough edges while keeping as much backward compatibility as is practical. The principal features of the updated XML format are the following:

- the updated CERIF XML allows for more compact mark-up;
- the whole CERIF XML mark-up shares one namespace;
- data of different CERIF entity types can be combined in one message;
- the XML elements representing multilingual attribute values can be embedded in the XML mark-up that corresponds to the base entity instance;
- the XML elements representing CERIF linking entities can be embedded in the mark-up of either end of the relationship;
- the old way of independent top-level only mark-up for multilingual and linking entities is also supported.

In this article we first explain the motivation for change, then we describe the updated CERIF XML exchange format, and outline some future steps. We assume readers to be familiar with the CERIF Entity-Relationship Model (Joerg et al. 2012).

2 Motivation for changing CERIF XML

The need for a new CERIF XML format has been motivated by experience gained in real world use cases. Both, the recent UK CRISPOOL (Clements and Lockhart 2010)[2] project and the Flemish FRIS[3] project revealed that the 2007 CERIF XML introduced a *fragmentation problem*, escalated by the fact that each entity in CERIF had its own namespace. Additionally, CRISPOOL demonstrated the impedance mismatch – a consequence being overly complex and highly

[2] JISC-funded CRISPool project - Final Report: http://www.st-andrews.ac.uk/crispool/media/crispool%20final%20report%20v2.1%20with%20appendices.pdf

[3] Flanders Research Information Space (FRIS): http://www.researchportal.be/

resource intensive transformation algorithms from the 2007 CERIF XML to CRIS systems – which had been discovered through usage.

2.1 CRISPOOL

CRISPOOL aimed at building a website to display research output from different research institutions. Each institution produced CERIF XML for relevant organisations, persons and research output, which was then uploaded to an aggregator (a Pure installation) to exhibit the information on the website. CRISPOOL clearly uncovered and demonstrated the fragmentation problem. Although the datasets being uploaded were not very large, the transformation from CERIF XML to the internal structure of the CRISPool CRIS was overly complex and very inefficient.

2.2 The FRIS Portal

With FRIS, a portal displaying research results and information (research output and projects) was built, and the research information uploaded in ER-inspired CERIF XML by the Flemish research institutions, where supplied datasets have been large. The FRIS project easily managed the fragmentation problem because its underlying data model was a clean CERIF relational model. However, future directions of FRIS aim to apply exchange models that fit better with current business needs - e.g. allow for more efficient data load times with the portal, and also with Web Service and Enterprise Service Bus technologies.

2.3 IST World

The EC-funded IST World project[4] was running under FP6 from April 2005 until November 2007, with partners from 14 European Countries. The objective of the project was to set up and populate an information portal with innovative functionalities to promote RTD competencies in IST – in the New Member States and Associate Candidate Countries – to facilitate and foster the involvement of different research entities in joint RTD activities. Advanced Visualization Techniques have been built on top of large data collections from more than 20 different sources. The project developed its own object-centric CERIF XML exchange format[5], because at that time, there was no CERIF exchange format available. The format has then been employed during all collection processes. Although then proposed and forwarded as a suggestion to the CERIF task group, back then, the ER-inspired CERIF XML (Joerg et. al. 2007) has been preferred.

2.4 Machine to Machine Exchange

Machine to machine exchange is often relevant for a modern CRIS e.g. with web-services, bulk transfer (import/export) between CRIS systems, or harvesting interfaces such as OAI-PMH

[4] Knowledge Base for RTD Competencies in IST (IST World): http://www.ist-world.org/

[5] IST World Formal Import/Export Specification: http://ist-world.dfki.de/downloads/deliverables/ISTWorld_D3.1_FormalImportExportSpecification.pdf

(Lagoze et al. 2002), where it is desirable, to process CERIF XML input and output in a stepwise manner – one conceptual entity in its entirety at a time (i.e. all information pertaining to one person, one organisation, or one publication, etc.). Hence, the input or output is treated as a stream of information to process one chunk of information and discard it, before processing the next. This is especially desirable when potentially dealing with very large data sets, but also supportive with compact responses from a web-service querying e.g. for a person and getting a CERIF XML representation of its conceptual whole – with all relevant information pertaining to the person as embedded elements. Finally, a single common type definition within one namespace only allows for the sharing of types – that is, one set of concepts, one set of restrictions, and one vocabulary. This vocabulary can be further reused and re-defined to form a model - i.e. the CERIF model in XML, and easily add extensions.

Both, the real world projects utilizing CERIF XML for their purpose but also the more generic scenario of machine-to-machine data exchange have contributed to the design of the updated CERIF XML format.

3 The updated CERIF XML

The updated CERIF XML overcomes the issues identified with the 2007 CERIF XML. We now briefly describe the main aspects of the update.

3.1 The single XML namespace

An important part of the update is the collocation of all CERIF-related XML mark-up in one namespace. In contrast with the former approach of defining separate namespaces per CERIF ER entities, the following advantages are seen:

- The XML mark-up of different CERIF entities is more easily combined in one XML message. This overcomes the fragmentation problem while being a precondition to another newly introduced XML construct: embedding.
- The grammar of the admissible CERIF XML mark-up is defined in one XML Schema. This allows for both a more thorough validation, as well as more efficient communication e.g. using Schema-aware XML-specific compression technologies such as EXI (Schneider and Kamiya 2011).

The XML namespace is identified by a single URI. It is constructed to contain the important information, while being compact. The current namespace URI is **`urn:xmlns:org:eurocris:cerif-1.4-0`** where the components are:

`urn`	- the URN scheme (as opposed to URL schemes);
`xmlns`	- the realm of XML namespaces;
`org:eurocris`	- the responsible organisation Internet identification (DNS domain name);
`cerif`	- the name of the product (CERIF);
`1.4`	- the release of the product (i.e. inline with the data model release);
`0`	- the release of the XML exchange format specification.

Future releases of the CERIF data model will have the product version incremented. Future improvements in the CERIF XML Schema generation process will be distinguished by an incremented exchange format version.

3.2 Mapping the CERIF ER Model into XML

The XML Schema for the updated CERIF XML is generated from the ER Model representation. This is easily possible due to its highly regular structure, namely the three basic kinds of CERIF entities: Research (Base, 2nd Level, other), Multilingual, and Link entities. An overview of the mapping for the different entities is given in Figure 1.

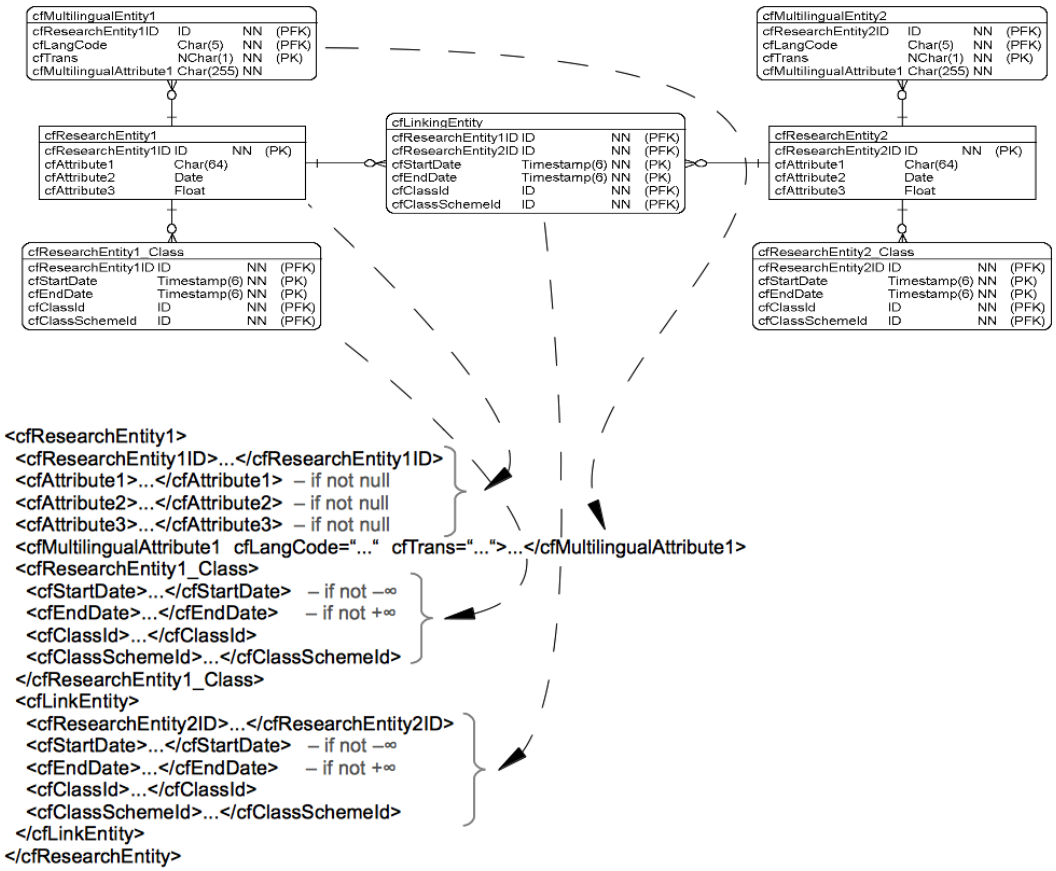

Figure 1: Transforming relational CERIF entities to embedding CERIF XML

We now explain each of the three basic CERIF entity kinds and their XML constructs' mapping, transformation or embedding.

Mapping CERIF Research Entities

Research entities are mapped 1:1. An entity instance is represented by an enclosing XML element, which embeds XML elements with non-null values from ER entity attributes.

```
<cfProj>
     <cfProjId>system-internal-project-identifier</cfProjId>
     <cfAcro>acronym</cfAcro>
     <!-- further attribute values and embedded contents here -->
</cfProj>
```

Transforming CERIF Multilingual Entities

Multilingual Entities are transformed to XML using a standardized construct: embedding as an XML element the multilingual attribute that contains the text value itself. This XML element has additional qualification attributes, namely (1) `cfLangCode` (the code of language and possibly also the country variant of the value), and (2) `cfTrans` (the translation mode).

```
<cfProj>
     <cfProjId>system-internal-project-identifier</cfProjId>
     <cfTitle cfLang="en" cfTrans="o">English original project title</cfTitle>
     <cfTitle cfLang="fr-CA" cfTrans="h">human translated French title</cfTitle>
</cfProj>
```

Transforming and Embedding CERIF Link Entities

Link Entities are transformed into XML in two possible ways: embedded under research entity 1, or embedded under research entity 2. Embedding subordinates the enclosing XML element to one of the two ends of the relationship. The identifier of the enclosing entity instance is then omitted: it is found one level higher in the XML.

```
<cfOrgUnit>
     <cfOrgUnitId>system-internal-organization-identifier</cfOrgUnitId>
     <cfName cfLang="en" cfTrans="o">organization name</cfTerm>
     <cfProj_OrgUnit>
          <cfProjId>system-internal-project-identifier</cfProjId>
          <cfClassId>classification-uuid</cfClassId>
          <cfClassSchemeId>classification-scheme-uuid</cfClassSchemeId>
          <cfStartDate>date-time when the relationship starts</cfStartDate>
     </cfProj_OrgUnit>
</cfOrgUnit>
```

Alternative embedding can be chosen:

```
<cfProj>
     <cfProjId>system-internal-project-identifier</cfProjId>
     <!-- … -->
     <cfProj_OrgUnit>
          <cfOrgUnitId>system-internal-organization-identifier</cfOrgUnitId>
          <cfClassId>classification-uuid</cfClassId>
          <cfClassSchemeId>classification-scheme-uuid</cfClassSchemeId>
          <cfEndDate>date-time when the relationship ends</cfEndDate>
     </cfProj_OrgUnit>
</cfProj>
```

Link Entities are embedded in Research Entities, and it is assumed that there exist records with system-internal identifiers. The updated CERIF XML only supports one degree of embedding; non-embedded objects in a linking relationship have to be referenced.

Unary Link Entities (such as `cfProj_Class`) can also be embedded under the base object:

```
<cfProj>
     <cfProjId>system-internal-project-identifier</cfProjId>
     <!-- … -->
     <cfProj_Class>
          <cfClassId>classification-uuid</cfClassId>
          <cfClassSchemeId>classification-scheme-uuid</cfClassSchemeId>
     </cfProj_Class>
</cfProj>
```

Mapping data types

Database types are approximated by XML Schema built-in types. As a rule, the XML Schema does not restrict the maximum lengths of character strings or size and precision of numbers. The only exception with mappings is the `ID` type (a dictionary type in the ER model): it limits the lengths of the identifiers of CERIF research entities to 128 characters.

Mapping valid time intervals

Valid time intervals of the relationships represented by the CERIF linking entities are conventionally expressed using the `cfStartDate` and `cfEndDate` ER attributes. These attributes are parts of primary keys and therefore cannot support null values. The fact that a relationship is not temporally bounded is conventionally represented by "sentinel" date-time constants:[6]

- a date-time value in `cfStartDate` that is sufficiently remote in the past represents "since time immemorial";
- a date-time value in `cfEndDate` that is sufficiently far in the future stands for "until things change".

Such a choice of constants plays sufficiently well with temporal bounds represented by real chronons (Özsoyoğlu and Snodgrass 1995) and makes querying the relational database viable, with one restriction: the sentinel values shall not be interpreted literally, but as replacements for the minus and plus infinities on the time axis. However, this should be considered a trick to make relational database technology work. As such, it should stay confined within the relational world.

The XML exchange format should rather reflect the real meaning of the data being exchanged. Outputting artificial values without giving consumers enough information to interpret them correctly would only lead to confusion. The right solution for the XML exchange format is to omit the corresponding mark-up altogether in the case of a temporally unbounded relationship, if there is no real `cfStartDate` and `cfEndDate` to represent in the first place.

Mapping financial amounts

A financial amount has two components: the currency unit and the number of these units. In the CERIF ER-model, these components are treated as inseparable, which also translates in the equivalent XML mark-up: ER attributes `cfAmount`, `cfPrice` or `cfTurn` are mapped to likewise called XML elements, while the accompanying `cfCurrCode` forms an XML attribute on the XML element. The value (the number of currency units) is expressed as a decimal number in the contents of the XML element.

---

[6] The constants are DBMS-specific.

For instance, the price of 9.95 € is expressed as:

```
<cfPrice cfCurrCode=”EUR”>9.95</cfPrice>
```

This construct is found both in CERIF research entities and in CERIF link entities.

3.3 Backward compatibility

An important design goal for the CERIF XML format update was to ensure that it accommodates all of the constructs from the original CERIF XML with minimum change. This was met: all pre-existing CERIF XML messages are valid with respect to the new schema, if their XML namespace is changed to the new one.

4 Towards CERIF 2.0 and Schema

With the uptake of the new CERIF XML format, further embedding candidates have been identified within the set of research entities. Increasingly, references to multiple standardized and open vocabularies are required. However, these are not necessarily exchanged in whole, but only their identifying references. The integration and thus validation of standardized vocabularies was an often-asked requirement and refers to the area of ontologies and linked open data. The newly identified requirements will be forwarded to the CERIF task group for further discussions and towards implementation with the next major release – CERIF 2.0.

5 Conclusion

An updated CERIF XML was constructed to overcome the rough edges of the 2007 CERIF XML. It provides the greater flexibility and scalability with well-defined object aggregations that were called for, while keeping backward compatibility with its predecessor and a firm connection to the CERIF data model and concepts. The updated CERIF XML is included in the CERIF 1.4 release, making the single extension from CERIF 1.3 and is provided under a CC-BY-ND license.[7] The backward compatibility allows existing CERIF XML producing agents to be used with just a minor adjustment. CERIF XML consuming agents will have to adapt to the greater variability of XML constructs that are now permissible. Early implementation efforts suggest that the required modifications should not be overly difficult. It is clear that an exchange format must support high flexibility towards underlying structures – but it must also be clear that a format is never final – and the next requirements are already being discussed in the CERIF task group.

---

[7] CERIF 1.4 by CERIF Task Group, euroCRIS is licensed under a Creative Commons Attribution-NoDerivs 3.0 Unported License. http://www.eurocris.org/CERIF-1.4/

Acknowledgements

We wish to thank the CRISPOOL and FRIS projects for their valuable input and feedback, and the CERIF task group for supportive discussions and directions. The work was partly supported also by EC Funding under META-NET with grant agreement no. 249119.

References

Clements A, Brander S, Heenan D, McCutcheon V, Brennan N, Brown J, Vestdam T (2012): *CERIF in Action: Synthesise, standardise and productionise CERIF for UK Higher Educational Institutions*, In Proceedings of CRIS 2012.

Castro J, Kolp M, Mylopoulos J (2002): Towards requirements-driven information systems engineering: the Tropos project. Information Systems 27(6), 365–389.

Fowler M (2003): *UML Distilled.* 3rd edition. Addison Wesley Professional, 2003. ISBN 9780321193681.

Hornbostel H (2006): *From CRIS to CRIS: Integration and Interoperability*. In Proc. 8th CRIS Conference, Leuven University Press, pp. 29—38, 2006.Ivanovic D. (2011): *Data Exchange between CRIS UNS, Institutional Repositories and Library Information Systems*. In Proc. of 5th International Quality Conference, 2011.

Jeffery K (2010): *The CERIF Model as the Core of a Research Organisation*. Data Science Journal Vol. 9, 2010.

Jeffery K, Asserson A (2010): Special Issue: *CRIS for European E-Infrastructure*. Data Science Journal, Vol. 9, 2010.

Jeffery K, Lay JO, Miquel JF, Zardan S, Naldi F, and Vannini Parenti I (1989): *IDEAS: A System for International Data Exchange and Access for Science Information Processing and Management* 25(6), pp. 703-711, 1989.

Jörg B, Jeffery K, Dvorák J, Houssos N, Asserson A, van Grootel G, Gartner R, Cox M, Rasmussen H, Vestdam T, Strijbosch L, Clements A, Brasse V, Zendulkova D, Höllrigl T, Valkovic L, Engfer A, Jägerhorn M, Mahey M, Brennan N, Sicilia M-A, Ruiz-Rube I, Baker D, Evans K, Price A and Zielinski M (2012): *CERIF 1.3 Full Data Model (FDM): Introduction and Specification,* euroCRIS, 2012.

Lagoze C, Van de Sompel H, Nelson M, Warner S (2002). *The Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0.*

Martin MD (1974): *Reference Manual for Machine-Readable Bibliographic Descriptions*, Unesco 1974, 71 p., compiler, Paris.

Özsoyoğlu G, Snodgrass TR: Temporal and Real-Time Databases: A Survey. In IEEE Transactions on Knowledge and Data Engineering, Vol. 7, No. 4, August 1995, pp. 513-532.

Schneider J, Kamiya T (editors): Efficient XML Interchange (EXI) Format 1.0. W3C Recommendation 10 March 2011. Available from http://www.w3.org/TR/2011/REC-exi-20110310/

van Godtsenhoven K, Karstensen ME, Sierman B, Bijsterbosch M, Hochstenbach P, Russell R and Vanderfeesten M (2009): *Emerging Standards for Enhanced Publications and Repository Technology: Survey on Technology.* Amsterdam University Press, SURF / EU Driver Series, 2009.

Wysocki and Tocatlian (1971). A World Science Information System: Necessary and Feasible. Taxon, Vol. 20, No. 4, August 1971, pp. 603 – 608. International Association for Plant Taxonomy.

Zimmermann E (2002): CRIS-Cross: Current Research Information Systems at a Crossroads. Simons E, Asserson A (eds), CRIS 2002, Leuven University Press, 2002.

Contact Information

Brigitte Jörg

German Research Center for Artificial Intelligence (DFKI GmbH)
Alt-Moabit 91c
D-10559 Berlin
Germany

brigitte.joerg@dfki.de

Jan Dvořák

Institute of Information Studies and Librarianship
Faculty of Arts
Charles University in Prague
U Kříže 8
CZ-15800 Praha 5
Czech Republic

jan.dvorak@ff.cuni.cz

Thomas Vestdam

Atira A/S
Nils Jernes Vej 10
DK-9220 Aalborg Oest
Denmark

tv@atira.dk